DOCUMENT RESUME

ED 381 563 TM 022 919

AUTHOR Schuyten, Shana; Tashakkori, Abbas
TITLE The Relationship between Assessor/Assessee Gender and Performance Observation Ratings.
PUB DATE Nov 94
NOTE 18p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Nashville, TN, November 9-11, 1994).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Beginning Teachers; Elementary Secondary Education; *Evaluators; Internship Programs; *Observation; *Performance; Pilot Projects; Public Schools; Sex Bias; *Sex Differences; Sex Stereotypes
IDENTIFIERS *Louisiana; Louisiana Teacher Internship Program; *Performance Based Evaluation

ABSTRACT

The effects of the genders of the assessor and the assessee on performance observation ratings of beginning teachers were studied in public schools in Louisiana. Data was collected in the pilot phase of the Louisiana Teacher Assessment Program for Interns, which included both teacher observation and structured interview. Of the assessees who reported their genders, 359 were female and 57 were male. Of assessors who reported gender, 468 were female and 195 were male. Dependent variables were the assessee's performance observation ratings. Independent variables were the genders of assessor and assessee. There were significant differences between the ratings of both male and female teachers, indicating that the gender of the assessee affected assessor ratings. However, no significant main effect of assessor gender was found in the results, suggesting that the bias of trained assessors does not seem to cause a major bias in the performance evaluation of teachers. Three tables present study data. (Contains 23 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 381 563

The Relationship Between Assessor/Assessee Gender and Performance Observation Ratings

Shana Schuyten
Louisiana Department of Education

and

Abbas Tashakkori
Louisiana State University

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

SHANA SCHUYTEN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Presented at the Annual Meeting of the MidSouth Educational Research Association, Nashville, 1994.

TM022919

BEST COPY AVAILABLE

As a profession, teaching is in the midst of dramatic reform. The impetus for this reform is the growing public discontent over the quality of education in the nation's schools (Tanner, 1993). As part of the reform, efforts are being undertaken to ensure that there will be a higher quality of education in the schools across the country. Widespread changes are being proposed regarding ways in which teachers are educated, trained, evaluated and certified. At the forefront of these changes are programs of induction and evaluation for beginning teachers. These programs have been developed for the enhancement and improvement of teaching in schools.

The attention to beginning teachers can be attributed to the research over the past two decades (e.g., Ryan, 1979; Tisher, 1978) that has shown that the first year of teaching is critical and often a difficult transition point in teacher development (Hoffman et al, 1986). In a comprehensive study of beginning teaching, McDonald (1980) reports:

> For most teachers, the initial experiences of teaching are traumatic events out of which they emerge defeated, depressed, constrained or with a sense of efficacy, confidence and growing sureness in teaching skills.(p.5)

McDonald also speaks of beginning teachers as being "abandoned" by the institutions where they receive their preservice training and are considered "peers" to all other teachers and by their employers. They have traditionally been left to their own devices and endure the first few years of teaching alone.

The reform movement, as it relates to beginning teachers, is being implemented primarily through policy initiatives at the state level. Few local school district policy makers and administrators are given the responsibility for devising and implementing methods of teacher evaluation. In 1980, there were only five district-supported programs

for beginning teachers, of which two were at developmental stages (Hoffman et al, 1986). A more recent survey indicated that more than 65% of all school districts in the United States have instituted some type of standardized teacher appraisal system (Katims & Henderson, 1990).

At the time of Hoffman's study (1980), only one state, Georgia, was active in the area of induction and evaluation of beginning teachers. Since that time numerous states have become active in this arena. Eight years ago a national survey of state activity in programs for beginning teachers identified 18 states with programs in advanced planning stages and 4 states with operational programs (Defino & Hoffman, 1986). Among those states were Georgia, Florida, Connecticut, Arizona, and Texas. All have mandated large-scale standardized teacher performance appraisal systems as part of efforts to reform and improve education in those states (Greenfield, 1987).

To investigate the phenomena of Competency Assessment, Sandefur (1983) conducted four annual surveys of the 50 states to provide data for analyzing nationwide trends in competency assessment of teachers. Ten years ago, the findings indicated that most state plans for teacher competency assessment included testing one or more areas of basic skills, professional or pedagogical skills, and academic knowledge. The testing was at the entry level, admission to the teacher education program, or prior to certification. At that time, a growing number of states had begun to require an internship or beginning teacher year with adequate assessment before initial certification was awarded. Sandefur's data analysis found that state competency assessment of programs had grown rapidly over the last six years and will continue to increase. He also indicated that continuing trends emphasizing testing in the basic skill areas will be used for

certification purposes. At that time, data also indicated that fewer states were using legislative action to mandate competency assessment of teachers; instead more states are relying on state department of education regulations.

Currently there is an increased demand for the identification of competent teachers within school systems across the nation. This demand, coupled with the availability of research and assessment instruments led to the development of large scale teacher assessment systems which were legislatively-enacted is such states as Arkansas, Connecticut, Florida, Georgia, Kentucky, Mississippi, Missouri, North Carolina, South Carolina, Tennessee, Texas and Virginia (Chauvin & Ellett, 1991; Ellett 1990). As many as eighteen states have utilized evaluation systems that were designed to include "on the job" assessment for purposes of teacher certification, merit pay, career ladders, and professional development (Association of Teacher Educators, 1988: Chauvin et al., 1991; United States Department of Education, 1987). If this trend continues, other states will follow their lead.

At the heart of these beginning teacher evaluation programs are classroom observation systems designed for certification or employment decisions. Any system which relies on observation as a mechanism for rating performance will be effected by the limits inherent in observational methods. While in general observational methods are advantageous due to the wealth of descriptive information they provide, there are some pronounced limitations as well.

Bias is the most obvious and limiting phenomena associated with observational techniques. Bias is an inclination or preference, especially one that interferes with impartial judgement; In effect, bias is prejudice (Webster, 1988).

Research indicates that observers are sometimes not very objective in their use of observational schedules. When objectivity is not maintained, the data the observer collects tend to reflect the biases and characteristics of the observer rather than the true performance that the observational measures sought to measure (Borg & Gall 1989). A recent review of evaluation literature shows that the differences among raters and observers are often attributable to differences in gender (Basow & Howe, 1987; Tieman & Rankin-Ullock, 1985). Differences attributable to gender may be the reason for existing stereotypical idealogies, differences in perceptions, attitudes, and opinions which may lead to observer bias. While other variables such as the observer's instructional level, teaching experience and highest degree attained have been considered, they are not nearly as poignant as the gender of the assessor as well as the assessee.

In a review of existing literature between 1932 and 1979, Feldman (1983) found that in seven studies, females were consistently rated higher than males. Female subjects also rated performance consistently higher than did male subjects (at least on some items) in studies by Basow and Distenfeld (1985), Basow and Howe (1987), Bennett (1982), and Harris (1976). In a 1989 study which studied the effects of gender, status and effective teaching on the evaluation of college instruction there was also some evidence of gender bias (Dukes & Victoria). In this study, male subjects rated female professors higher than female subjects. In all of these studies, it was found that the sex of the rater and ratee interfered with the objectivity and credibility of the rater, as well as the validity of the ratings.

In other gender research, such as Levenson, Bufford, Bonnoe and Davis (1975) and Tieman and Rankin-Ullock (1985) male subjects rated performance higher than female subjects. Goldberg's 1986 study was ideally suited to the examination of teaching evaluations and results also uncovered a bias in which subjects continually favored a male over a female. Subsequent research showed that males were evaluated higher than females for the same performance and that the status of the person being evaluated altered the situation. It is clear that the literature has produced mixed findings about the relationship between the rater and ratee's gender and its effect on the evaluation of effective teaching.

The purpose of the present study is to determine the effects of assessor and assessee gender on performance observation ratings. We seek answers to 3 questions in this respect: 1) Do male and female assessors differ in their average ratings of assesses?; 2) Is there a difference in the average ratings of male and female assessees?; and 3) Is there a same-sex or cross-sex bias in performance ratings of intern teachers' effectiveness?

The data obtained in the evaluation of a group of beginning teacher interns in public schools in Louisiana were used to answer these questions. The data was collected at the pilot phase of the *Louisiana Teacher Assessment Program for Interns*. The program composed of two primary data collection methods: classroom observation and structured interview. Each intern teacher was assigned to a team of three highly competent, experienced educators **(assessors)** with each conducting a minimum of one visit to the intern's classroom during each semester of the year. During these visits each assessor conducted the evaluation utilizing the **Louisiana**

Teacher Assessment Instrument (LTAI). The LTAI pilot test used a three-point rating scale designed to allow formative feedback to the intern teacher. The three points are defined as 1) Needs improvement, 2) Proficient, and 3) Commendable.

Methods

Sample

The target population for the 1993-94 Pilot Study included all interns teaching in Louisiana's public schools. In selecting the sample for the Pilot Test, **local educational agencies (LEAs)** were identified based on the projected number of interns (Bulletin 1472, 1991-1992 Annual Financial and Statistical Report), representation from each of the eight Regional Service Centers, geographic proximity for assessor training and intern orientation, gender, and ethnicity. If a LEA agreed to participate, all interns in the LEA were included in the Pilot Test. Originally, 13 LEAs agreed to participate. However, in these districts the actual number of interns was lower than projected. To augment the sample, four additional LEAs agreed to participate, therefore, all the interns from these four LEAs were included. Furthermore, interns were sampled from three other LEAs. Interns from 20 LEAs participated in the Pilot Test.

Existing data represent assessment ratings collected during the 1993-1994 Pilot Test. Of the assessees who reported their gender 339 (85.6%) were female and 57 (14.4%) were male. Of the assessors who reported their gender 468 (70.6%) were female and 195 (29.4%) were male. The percentages of male and female in the sample of assessees (interns) approximate the respective percentages of Louisiana's

public school teacher population. Data from the LTAI were available for 410 of the 430 interns and for all of the 721 assessors. Table 1 shows the number of observations in each of the 4 groups under study.

*** Table 1 about here ***

Variables

The independent variables for this study are the gender of both the assessor and assessee. Gender is a self-reported assigned variable which both the assessee and the assessor report on their demographic data form prior to assessment.

The dependent variables are the assessee's performance observation ratings. The assessment ratings collected during the 1993-1994 Pilot Test are available at the rater (assessor) level. For each assessee, there are performance observation ratings from each of three assessors across 27 attributes. All attributes are rated using a three-point scale. These 27 attributes are expected to represent 8 components of effective teaching. For the present study, for each "component" of effective teaching, a composite score was constructed by calculating the sum of the attribute ratings which were expected to represent the component.

Results

Table 2 presents the mean of the 8 components of effective teaching for the 4 groups under study. It should be noted that since the number of items were not the same for all components, the means are not comparable across components. However, each performance rating mean is comparable across the 4 groups. As Table 2 indicates, on the average, the component ratings were consistently higher for

female assessees, as compared to the male assessees. This trend was present regardless of the assessor gender. However, when a male assessor rated a female assessee, higher ratings resulted than when a female assessor rated a female assessee. Male assessors gave female assessee higher ratings on 5 of the 8 components (IIA, IIB, IIC, IIIB, IIIC). Female assessors gave female assessees higher ratings on only 3 of the 8 components (IA, IIIA, IIID). It is apparent that female assessees are rated highest, when assessed by male assessors.

*** Table 2 about here ***

As Table 2 also indicates, among the male assessees, those who were rated by female assessors scored higher than those rated by male assessors. The difference in mean component ratings among those males assessed by females and those assessed by males is similar to the difference across the sample of male and female assessees.

In line with these findings, a 2 x 2 (Assessor Gender by Assessee Gender) Multivariate Analysis of Variance (MANOVA) was performed with the 8 component composite scores as dependent variables. MANOVA indicated a statistically significant main effect of assessee gender ($F(8,1037)=4.99$; $p<.001$). The main effect of assessor gender, and the interaction of the two independent variables were not statistically significant.

Following the significant multivariate main effect of assessee gender, univariate analysis of variance was performed on each of the 8 dependent variables. Results indicated significant differences between the male and female assessees in 6 of the 8 components. Table 3 presents the results.

*** Table 3 about here **

In the above-mentioned analyses, the number of observations were not the same across the 4 groups of assessors and assessee gender. A regression approach was used in the calculation of sum of squares. As a precaution, a second analysis was also performed after subsamples were selected such that the 4 groups had equal numbers (79 female and male assessees and 79 female and male assessors). MANOVA lead to the exact pattern of results as in the full sample, with only the main effect assessee gender being significant ($F(8, 286) = 2.67$; p.05).

Discussion

In effect a performance evaluation is a type of social perception. As such, it inevitably entails "forming beliefs about the quality of a person's task performance based upon perceptions of the person's activities (Foschi & Lawler, 1994)". When performance evaluations are not structured in a way that successful and unsuccessful outcomes are distinct and easy to judge, evaluation may be difficult. When this is the case, perceptual biases such as those mentioned may come into play. These can have a powerful effect on performance evaluations. Perceptual biases involving social characteristics such as gender might ultimately bias the assessment process.

It is important that we examine the possibility of any gender biases that may devalue or discredit the significance of assessment results. As the nation moves toward teacher assessment systems that rely on observational performance ratings, we must be prepared to extrapolate true assessment ratings from those that are confounded by gender bias. Differences in assessment results are tolerable, but not if

they are the result of gender biases rather than true differences in the assessees performance.

A prediction of the study was that male and female teachers would be evaluated differently by assessors. As expected, there were significant differences between the ratings of both male and female teachers. In other words, the gender of the assessee effected the ratings that the assessor gave. Although the effect was present for both sexes, these differences were greater for male assessors than for female assessors.

Another prediction of the study was that male and female assessors would be different in their average ratings of the assessees. It is comforting that in line with a weak trend in the results in support of this prediction, statistical tests did not support it. No significant main effect of assessor gender was found in the results. At least in the present context, the gender of trained assessors does not seem to cause a major bias in the performance evaluation of teachers.

It was also predicted that evaluation of male and female teachers by same-sex assessors would be different from opposite-sex assessors. As the results indicate, female assessees scored higher on 5 of the 8 components when assessed by a male assessor rather than a female assessor. Also, that male assessees were rated more positively by female assessors than when rated by male assessors was in favor of the interaction hypotheses. However, in the absence of a significant interaction effect, these results should only be interpreted as tentative.

The slightly greater ratings of female teacher interns might be an indication of their greater effectiveness as teachers. However, when the assessors were female, the difference between the two assessee genders were smaller than when the assessors were male. This is inconsistent with the assumption of greater effectiveness among females, and might point to a slight bias in evaluations. The magnitude of such bias, if it exists, is relatively small, possibly due to the fact that the raters were highly trained for the task.

Table 1. Number and Percentage of Observations in each of the 4 Groups

	Assessee Gender	
Assessor Gender	**Female**	**Male**
Female	**617** (68.1%)	**79** (54.5%)
Male	**289** (31.9%)	**66** (45.5%)
Total	**906** (86.2%)	**145** (13.8%)

Table 2. The Mean Ratings of the 8 Components in each of the 4 Groups

Assessor Gender	Male		Female	
Assessee Gender	Male	Female	Male	Female
	X SD	X SD	X SD	X SD
IA	13.65 2.33	14.49 2.59	13.89 2.45	14.91 2.59
IIA	4.92 .966	5.33 .838	5.06 .952	5.30 .904
IIB	4.77 1.10	5.04 .954	4.96 1.09	4.96 1.10
IIC	4.71 1 .14	5.10 .984	4.75 1.05	4.97 1.08
IIIA	9.12 1.75	9.92 1.79	9.40 1.81	9.93 1.91
IIIB	9.78 1.71	10.17 1.67	9.97 1.66	10.02 1.85
IIIC	8.83 1.83	9.76 1.69	9.14 1.92	9.68 1.83
IIID	11.66 1.98	12.38 2.14	11.60 2.08	12.52 2.15
SAMPLE TOTAL	8.43	9.02	8.60	9.04

Dependent Variables (Components)

IA= Teacher Plans Effectively For Instruction
IIA= Teacher Maintains Enviornment
IIB= Teacher Maximizes Time
IIC= Teacher Manages Learner Behavior
IIIA= Teacher Delivers Instruction
IIIB= Teacher Presents Content
IIIC= Teacher Provides For Student Involvement
IIID= Teacher Assesses Student Progress

Table 3. Results of Anaylsis of Variance.

F Values for the Main-Effects and Interaction

	MANOVA		**Univariate ANOVA***							
Source	**F**		**IA**	**IIA**	**IIB**	**IIC**	**IIIA**	**IIIB**	**IIIC**	**IIID**
Assessee Gender	**4.99***	**F**	**15.98***	**15.8***	2.04	**9.73***	**15.44***	1.88	**20.20***	**17.73***
Assessor Gender	.675	**F**	1.97	.473	.285	.176	.805	0.016	.526	.044
Assessee Gender X Assessor Gender	1.09	**F**	.151	1.05	2.02	.831	.647	1.12	1.35	2.65

***p < .001**

Dependent Variables (Components)

IA= Teacher Plans Effectively For Instruction
IIA= Teacher Maintains Enviornment
IIB= Teacher Maximizes Time
IIC= Teacher Manages Learner Behavior
IIIA= Teacher Delivers Instruction
IIIB= Teacher Presents Content
IIIC= Teacher Provides For Student Involvement
IIID= Teacher Assesses Student Progress

REFERENCES

Association of Teacher Educators. (1988). Teacher Assessment. Reston, VA: Author. (ERIC Document Reproduction Service No. ED 289 869)

Bascow, S.A., & Distenfield, S. (1985). Teacher expressiveness: More important for male teachers than female teachers. Journal of Educational Psychology. 77, 45-52.

Bascow, S.A., & Howe, K.G. (1987). Evaluations college professors: Effects of professors' sex-type and sex, and student's sex. Psychological Reports. 60, 671-678.

Bennett, S.K. (1982). Student perceptions and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluation. Journal of Educational Psychology. 74, 170-179.

Borg, W., & Gall, M. (1989). Educational Research an introduction (3rd ed.) White Plains, NY: Longman.

Chauvin, S.W., Loup, K.S.., & Ellett, C.D. (1991, April). Development and validation of a comprehensive assessment system for teaching and learning. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago. (ERIC Document Reproduction Service No. ED 335410).

Defino, M., & Hoffman, J. (1986). A status report and content analysis of state mandated teacher induction programs. (Technical Report No. 9057) Austin, TX: The University of Texas at Austin, Research and Development Center for Teacher Education.

Feldman, K.A. (1983). Course characteristics and college teachers as related to evaluations they receive from students. Research in Higher Education. 18, 3-124.

Foschi, M., & Lawler, E. (1994). Group Processes: Sociological Analyses. Chicago: Nelson-Hall Publishers.

Goldberg, P.A. (1968). Are women predjudiced against women. Transaction. 5:28-30.

Greenfield, William (1987). Instructional Leadership. Boston: Allyn and Bacon, Inc.

Harris, M.B. (1976). The effects of sex, sex-stereotyped descriptions and institution on evaluations of teachers. Sex Roles 2:15-21.

Hoffman, J. V., Griffin, G.A., Edwards, S.A., Paulissen, M. O., O'Neal, S.F., & Barnes, S. (1986). Teacher Induction Study: A final report of a descriptive study. Austin, TX:The University of Texas at Austin, Research and Developmental Center for Teacher Education.

Katims, D. & Henderson, R. (1990). Teacher Evaluation in Special Education. NASSP-Bulletin, 54(527), 47-52.

Levenson, H., B. Buford, B. Bonno, and L. Davis. (1975). Are women still prejudiced against women? A replication and extension of Goldberg's study. Journal of Psychology 89:67-71.

Louisiana State Department of Education. (1993).Louisiana Assessor Training Manual: Intern Teachers. (LDE R.S. 17:3721). Baton Rouge, LA: LDE Printing Office.

McDonald, F. (1980). The problems of beginning teachers: A crisis in training (Vol 1). Study of induction programs for beginning teachers. Princeton, NJ:Education Testing Service.

Ryan, K. (1979). Toward understanding the problem: At the threshold of the profession. In K. Hokey and R. Bents (Eds.), Toward meeting the needs of beginning teachers. Minneapolis, MN: United States Department of Education/Teacher Corps.

Sandefur, J.T. (1983). Competency Assessment of Teachers: 1980-1983. New York: Macmillian.

Schwab, R.L. (1991). Research-Based Teacher Evaluation. Boston: Kluwer Academic Publishers.

Tieman, C.R., and Rankin-Ullock. (1985). Student evaluations of teachers. Teaching Sociology 12: 177-191.

Tisher, R. (Ed.).(1978). The induction of beginning teachers in Australia. Melbourne, Australia: Monash University.

United States Department of Education. (1987). What's happening in teacher testing. Ananalysis of state teacher testing practices. Washington, DC: Author.